Combining Multiple Scoring Systems For Video Target Tracking Based on Rank-Score Function Variation

D. Frank Hsu and Damian M. Lyons

Robotics & Computer Vision Laboratory
Department of Computer & Information Science
Fordham University
Bronx NY 10458
{hsu,lyons}@cis.fordham.edu

Keywords: Feature selection, Multiple scoring system, Rank combination, Rank-Score function, Rank-Score graph, Score combination, Target tracking

Abstract

Tracking of video targets is the process of estimating the current and predicting the future state of a target from a sequence of video sensor measurements. Multitarget video tracking is complicated by the fact that targets can occlude one another and affect video feature measurements in a highly non-linear and difficult to model fashion., Tracking multiple targets that undergo repeated mutual occlusions is a challenging problem with several issues to be addressed. In this paper we propose a multisensory fusion approach to the problem of multitarget video tracking with occlusion. Each sensory cue is treated as a scoring system on the set of possible target tracks. Scoring behavior is characterized by a rank-score function, defined by Hsu and Taksa [11]. A diversity measure defined by Hsu, Chung and Kristal [7] is used based on the variation in rank-score functions. We describe the importance of using the rank-score function in the combination of multiple scoring systems for tracking multiple targets with repeated target occlusion, in particular in the process of hypothesis pruning and feature selection. We present experimental results for 12 video sequences from a variety of situations that demonstrate that our approach can be used to design a feature and fusion selection criterion that improves video tracking performance for situations with multiple, mutually occluding targets.

1. Introduction

Tracking of targets automatically in video is a problem with a number of applications, including automated surveillance, robotics and virtual reality, amongst others. However, it remains a difficult problem, especially when handling video with multiple targets and crowded scenes [6]. Unfortunately, a video camera looking at an airport lobby or a busy city intersection will have exactly this kind of scene, and this motivates our interest in finding an approach to tracking that works well in such cases.

The image of a video (or information from a sensor) can be a very rich source of information about a target: image position, image velocity, color properties, shape properties and so forth. Fusing multiple sources of sensory information therefore is an appealing way to make tracking more robust [24]. Existing approaches to sensory fusion for video tracking have tended to fall

into one of three categories: statistical approaches, physical modeling approaches and heuristic approaches. In this paper we propose a new approach based on the emerging field of Combinatorial Fusion Analysis (CFA) (see [7]) which has been applied to other fields such as information retrieval, pattern recognition, virtual screening and drug discovery, and protein structure prediction ([8],[11],[15] and [26]).

Much work has been done in fusion for multisensory video tracking in the past. They can be divided into three categories: statistical, physical and heuristic. The first category, and arguably the largest, represents the sensory measurements as random variables whose probability density functions can be characterized and used to define a sensory fusion operation. The target tracking community has developed a number of such elegant approaches [1]. These include the Kalman filter, Reid's Multiple Hypothesis Tracking (MHT) algorithms [20], the result by Cox and Hingorami [3], Sharma's [22], Maximum Likelihood (ML) and Maximum A Posterior (MAP) formulations, and the results of Rasmussen and Hager [19]. If the image generation process can be modeled in sufficient detail, then this physical model can be used to determine how sensory measurements should be fused. This gives rise to the second category of work in fusion: physical modeling by Nandhakumar and Aggarwal [17]. The final category of work is the heuristic category. The fusion of data in this case is based on a proposed heuristic measurement, derived from a pragmatic appreciation of the nature of the problem. These results include Checka and Wilson [2], Loy et al [14], and Snidaro et al [23].

The statistical and physical modeling approaches both rely on being able to model correctly and efficiently the relationship between feature measurements and target state. However, when one or more targets engage in repeated mutual occlusions, the relationship between targets and feature measurements can become highly non-linear and very difficult to model. The heuristic approaches sidestep this problem by adopting an approximate, rather than exact statistical or physical model. The disadvantage is, however, that there is no guarantee of performance.

The work described in this paper follows a new approach to the problem of fusion for video tracking that can encompass the three categories described. It does not assume a single statistical or a physical modeling approach. It does not propose a specific heuristic rule or architecture for fusion. Instead, our approach relies on the measurement data themselves and then use the diversity between these measurements to determine the best fusion rule. Our principal tool is the emerging field of Combinatorial Fusion Analysis (or CFA) [7]-[8], [10], [11] -[15]). The use of CFA has several distinct characteristics which distinguish it from existing fusion approaches (see e.g., [24]-[25]). It is bottom-up and does not impose a model on the measurements. It begins by incorporating a variety of *scoring systems* obtained at the data level or the decision level. At the data level, these multiple scoring systems represent different sensory cues, features, or combinations of cues and/or features. At the decision level, they are derived from methods such as statistics, physical modeling, heuristics, analysis, combinatorics and computation. We refer readers to the book chapter [7] for a survey, summary and description of CFA. In this paper, we mainly use CFA at the data level to represent cues or features as multiple scoring systems. In our approach, we consider: (a) both score and rank function for each feature or piece of evidence to be combined, and explores the interaction between the two functions, and (b) both combinatorial possibility and computational efficiency for combining multiple scoring systems.

In Section 2, we present a fusion framework for tracking applications that supports the evaluation of the combinatorial options for fusion; selection of fusion operation as well as the

selection of which subset of cues to fuse. Section 3 describes target hypothesis pruning and feature selection. Experimental results are presented in Section 4. The experiments were performed on twelve video sequences containing a range of tracking situations including several with multiple mutual occlusions. The experimental results show that CFA can be used to design a fusion selection criterion for a range of features and fusion operations that produces a significant improvement in tracking accuracy. Section 5 presents our conclusions and future plans.

2. Combination of Multiple Scoring Systems

In our work [10], [16], we have proposed a multiple hypothesis framework for implementing and evaluating a variety of feature fusion operations, including score and rank fusion combinations, for video tracking applications. In this paper, we revise and update that framework and use it as a basis for our experimentation.

2.1 Score and Rank Functions of a Scoring System Module.

Let $\boldsymbol{D}_j = \{d_1,...,d_n\} \subseteq \mathbf{D}$ be the labels of track hypotheses in the pool of n track hypotheses for target $j \in 1, ..,p$ generated by the collection of scoring systems. We will assume that each module operates on the same pool of track hypotheses. This could be by use of a common hypothesis generation stage [16], as in our case, or by the generation of a set of composite tracks [18].
The score function $s_{kj}(d)$ assigns a real number to each d in $\boldsymbol{D}_j$ which is the score given by the tracking module M_k to the candidates for the j^{th} target. When treating $s_{kj}(d)$ as an array of real numbers, it would lead to a rank function $r_{kj}(d)$ after sorting the $s_{kj}(d)$ array into descending order and assigning a rank (a positive natural number) to each of the d in $\boldsymbol{D}_j$. The resulting rank function $r_{kj}(d)$is a function from $\boldsymbol{D}_j$ to $\boldsymbol{N}=\{1,2,\ldots,n\}$ (we note that $|\boldsymbol{D}_j|=n$).

In order to properly compare and correctly combine score functions from multiple scoring systems (multiple features for a single sensor, or multiple items of evidence from multiple sensors) normalization is needed. We simply adopt the following transformation from $s_{kj}(d):\boldsymbol{D}\rightarrow\boldsymbol{R}$ to $s^*_{kj}(d):\boldsymbol{D}\rightarrow[0,1]$ where $s^*_{kj}(d) = \dfrac{s_{kj}(d)-s_{\min}}{s_{\max}-s_{\min}}$, $d \in \boldsymbol{D}$ and $s_{max}=max\{ s_{kj}(d)| d \in \boldsymbol{D}\}$ and $s_{min}=min\{ s_{kj}(d)| d \in \boldsymbol{D}\}$.

2.2. Rank and Score Combinations

When given m scoring systems for a target j with score functions $s_{kj}(d)$ and rank function $r_{kj}(d)$ and $k=1,2,\ldots,m$, there exist several different ways of combining the output of the scoring systems, including score combination, rank combination, voting, average combination and weighted combination. For the m scoring systems with $s_{kj}(d)$ and $r_{kj}(d)$, we define the score functions s_R and s_S of the rank combination (RC) and score combination (SC) respectively as: $s_R(d) = \sum_{k=1}^{m}\left[w_k r_{kj}(d)\right]$, and $s_S(d) = \sum_{k=1}^{m}\left[v_k s_{kj}(d)\right]$. For this paper, we will define $w_k = 1/m$ and $v_k = \dfrac{1/\sigma_k^2}{\sum_1^m 1/\sigma_k^2}$ where σ_k^2 is the variance in s_{kj}. That is, the rank combination is an average rank combination, and the score combination is a Mahalanobis combination. We choose this score

combination because of its connection to the Bayesian formulation, and its widespread use in tracking.

As we did before, $s_R(d)$ and $s_S(d)$ are sorted into ascending and descending order to obtain the rank function of the rank combination $r_R(d)$ and the score combination $r_R(d)$, respectively.

2.3. The Rank-Score Graph of a Scoring System Module

Recently, Hsu and Taksa [11] characterize the relationship that an expert habitually produces between score and rank as the *rank-score function*s and the graph of that function as the *rank-score graph* (Fig. 2.1)*;* the graph of a monotonic function *f* that relates the rank and score of a set of candidates. Let $s : \boldsymbol{D} \rightarrow \boldsymbol{R}$, where $s(d)$ is the score of candidate d in the set of candidates $\boldsymbol{D}$. Let $r : \boldsymbol{D} \rightarrow \boldsymbol{N}$, where $r(d)$ is the rank of candidate d when the candidates are ordered according to their score. Then, *the rank-score function f* is the composite of s and r defined as $f : \boldsymbol{N} \rightarrow \boldsymbol{R}$, where

$$f(i) = (s \circ r^{-1})(i) = s(r^{-1}(i)).$$

By our definition of rank, a rank-score function has to be *monotonic non-increasing.* However, the shape of the graph can be different for different experts and is a characteristic of that expert's approach. So, the expert who assigns scores in a linearly decreasing fashion will have a linear rank-score graph (e.g., Fig. 2.1 (f_2)). The expert who habitually assigns high scores to a large subset of its top ranked candidates will have a graph that is not a straight line, but has a low slope for the top ranked candidates and a higher slope for the remainder. The concave-down graph f_3 in Fig. 2.1 is an example of this. A third class of scoring behavior is exemplified by f_1 in Fig 2.1. In this case, the expert habitually gives higher scores to a small subset of its top ranked candidates and much lower scores to the rest.

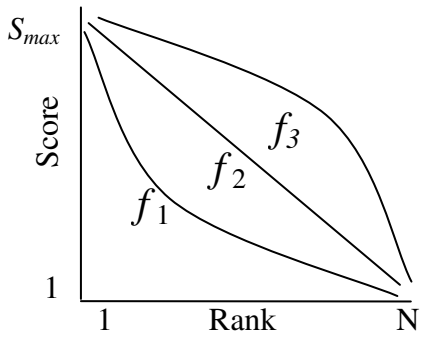


Figure 2.1: Rank-Score Graphs

Hsu and Taksa [11] indicate that a diversity measure based on the rank-score graph can be used to determine whether a score or rank fusion will produce a better result. Hsu and colleagues have used the new paradigm for diversity measurements between two scoring systems in a variety of applications, including information retrieval ([11]), protein structure prediction ([13]), target tracking ([8], [10] and [16]) and combinatorial fusion ([7] and references). When the rank-score graphs of two experts are very similar, then a score combination will produce the best fusion. When the rank-score graphs are very different, then a rank combination produces the better result.

2.4. Diversity between Scoring System Module Characteristics.

Returning to the rank and score function definitions of Section 2.1, it is now possible to define a set of rank-score functions. The rank score function for tracker module k for target j is:

$$f_{kj} : \boldsymbol{N} \rightarrow \boldsymbol{R},\ f_{kj}(i) = s_{kj}(r_{kj}^{-1}(i)) = \text{score of track hypothesis } d \in D_j \text{ which has rank } i$$

The rank-score graph of the scoring system module k for target j is the graph of the rank-score function f_{kj}.

Different tracking modules will be affected to differing degrees. Hsu and Taksa's results [11] indicated that when the graphs for a target become sufficiently different, a rank fusion operation will most likely perform better than a score fusion operation. Intuitively, this means that the scoring behavior of the tracking modules have become sufficiently different that a numerical combination would be biased. Thus, the rank-score diversity could be a useful criterion for deciding whether to use a rank or a score fusion operation.

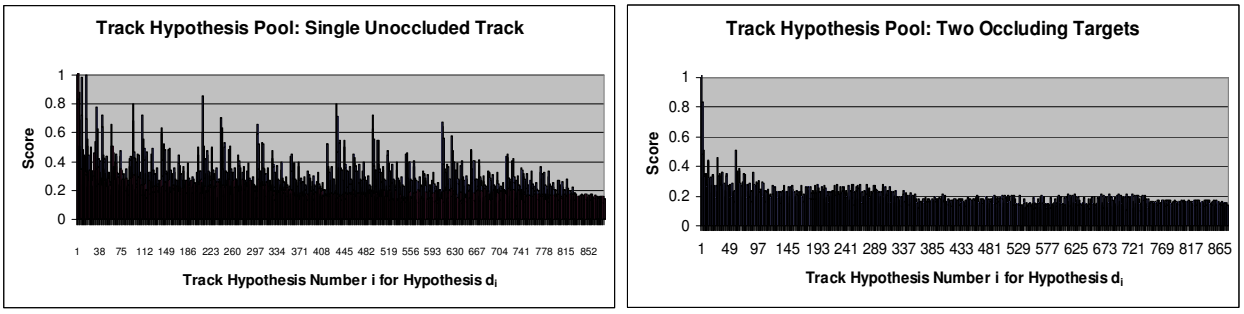
We compare the rank-score graphs from each scoring system module for each target to determine which to use, and which fusion operation to employ. We define the difference between two rank score graphs f_A and f_B as follows:

$$d(f_A, f_B) = \sum_{i \in N} (f_A(i) - f_B(i))$$

The results of Hsu and Taksa's ([11]) indicate that for two modules *A* and *B*, when $d(f_A, f_B)$ is sufficiently large, then rank fusion will outperform score fusion for these two modules *A* and *B*. In Section 4 we evaluate this proposition experimentally by looking at the combinatorial combinations of the fusion operations and evaluating the relationship between this diversity measure and a ground-truth based performance measurement. The results of this study will demonstrate that this diversity measure is a useful criterion for selecting fusion operations.

3. Target Hypothesis Pruning and Feature Selection

We describe the importance of using the rank-score function in the combination of multiple scoring systems for tracking in a scenario with repeated mutual target occlusion. In particular we compare this heavy occlusion scenario with a much simpler, unoccluded tracking scenario for two tasks important for feature combination in tracking: (a) target hypothesis pruning, and (b) feature selection. We have shown that in the heavy occlusion scenario, using rank and score combination has distinct advantages in target hypothesis pruning (Hsu and Lyons [9], Hsu et al. [10]). On the other hand, we have also shown that the rank-score function and the variation of the rank-score function among individual scoring systems can be used to select features which improve the rate of false positives (FP) and false negatives (FN) of the combined scoring system (Hsu and Lyons [9]).



(a) Single, Unoccluded Target (b) Two Partially Occluding Targets

Figure 3.1: Typical Score Distributions for Two Tracking Scenarios

Hsu and Lyons [9] explored some of the theoretical implications of rank versus score in tracking. Figure 3.1 below shows examples of typical track hypothesis score distributions for two tracking scenarios. The graph data were collected with the RAF tracker [10] using the position tracking feature module (tracking the location of the centroid of each foreground region). The track hypothesis pool was logged in each case after tracking had proceeded for approximately 15 frames. The distribution in Fig. 3.1(a) was produced by tracking a single, unoccluded target, a

person walking. The distribution in Fig. 3.1(b) was the result of tracking two targets that engaged in repeated mutual occlusions, two people walking as a couple. Hsu and Lyons note that for these two cases of typical tracking scenarios, the single target tracking scenario produces a greater variance in scores, because the scoring system can distinguish good target hypotheses. For the crowded tracking scenario, there is less variance observed, because the scoring system has difficulty distinguishing good and bad hypotheses; the correct choice of target is less clear cut.

3.1 Target Hypothesis Pruning:

Hsu and Lyons [9] propose that the track hypothesis probability distributions are different in these cases. They propose the track hypothesis probability histograms h_a and h_b in Fig. 3.2 as typical for scenarios such as those in Fig 3.1(a) and Fig. 3.1(b) respectively (where the vertical ϕ-axis is frequency and the horizontal p-axis is probability of a track hypothesis). By definition of h_a we note that approximately $\int_0^1 h_a(p)dp = \int_0^1 h_b(p)dp$. Let p_c be the value of p when h_a and h_b intersect. In our example graphs, p_c is close to 0.5. The histogram h_a reflects that there are fewer hypotheses with good scores (to the right of p_c) than other hypotheses with clearly worse scores (to the left of p_c). On the other hand, h_b has similar numbers of hypotheses with good and bad scores. Based on this proposition, they then show pruning the pool of tracking hypothesis **D** has a much greater effect on the variation in ranks in a crowded tracking scenario (Fig. 3.2 (b)) than in a sparse tracking scenario (Fig. 3.2 (a)).

The graphs in Fig 3.2 are used to derive the rank-score graph associated with the scoring system for each of the two scenarios. The rank of a track hypothesis is related to its score (probability) and the score (probability) histogram as follows:

$$r(s) = \sum_{x=s}^{1.0} h(x) \cong \int_s^1 h(x)dx$$

The rank functions r_a and r_b for h_a and h_b respectively in Fig. 3.2 can be derived in this fashion and graphed against score to yield the rank-score graphs f_a and f_b in Fig. 3.3. From the rank-score graphs in Fig. 3.3 it can be shown (see [9] for details) that if a probability cutoff p_x is used to prune the track hypothesis pool, then as long as $p_x > p_c$ this will produce a greater variation in

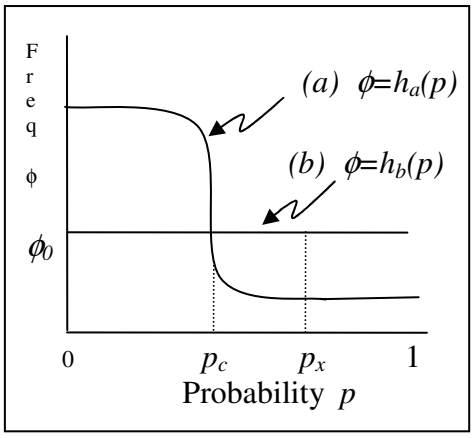


Figure 3.2: Frequency of Probabilities for Track Hypotheses in (a) Sparse Scenario, and (b) Crowded Scenario

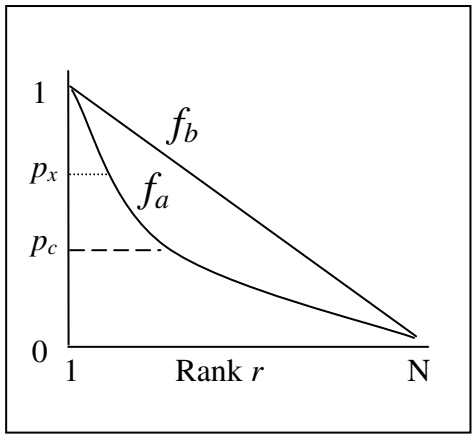


Figure 3.3: Rank-Score Graphs f_a and f_b derived from probability histograms h_a and h_b respectively

ranks in the crowded scenario (f_b in Fig. 3.3) than in the sparse scenario (f_a in Fig. 3.3). This is apparent from Fig. 3.3 since f_a has a steeper slope than f_b in the interval $p_x > p_c$ and $f_b^{-1}(p_c) - f_b^{-1}(p_x) > f_a^{-1}(p_c) - f_a^{-1}(p_x)$. As previously mentioned, Hsu and Taksa [11] shows that the variation of the graph of the rank-score function between two experts has impact on whether a rank combination or score combination produces a better result. Hence, these results illustrated that for a crowded scene, the benefit of score based fusions and of rank-based fusions will vary depending on the hypothesis pool pruning threshold. This explains why in crowded tracking scenarios, working with rank and score combinations has distinct advantages, because the same score probability cutoff p_x produces more variations in rank in f_b than that in f_a. So in that case, working directly with rank combinations can produce a more accurate result.

3.2 Feature Selection:

In Hsu and Lyons [9], they also examine the implication of rank versus score in selecting features for fusion when tracking in a crowded scenario. That is, restricting our attention to Fig 3.2(b) but considering *more* than one scoring system. They note that f_b in Fig. 3.3 is the typical form of the rank-score graph in this case as related to $h_b(p)$ in Fig. 3.2 with the tracking scenario for occluded targets in Fig 3.1(b) . However, a given scoring system will vary from this typical case, and may produce a rank-score graph that curves above or below this 'ideal' case. This is shown in Fig. 3.4, where h_{b1} and h_{b2} are the histograms for two different scoring systems when tracking in the crowded scenario. We note again that since $\int_0^1 h_{b1}(p)dp = \int_0^1 h_{b2}(p)dp$ approximately, the up-down curve properties of h_{b1} and h_{b2} have to be opposite. This leads to the rank-score graph of f_{b1} and f_{b2} respectively in Fig. 3.5.

The feature selection problem can be phrased as: given the scores for each hypothesis for each feature, which features should be fused to produce the best performing result. Lyons and Hsu [9] use the number of false positives (FP) and false negatives (FN) associated with the combination as their criterion for evaluating performance. They show that if scoring systems with *complementary* rank-score functions f_1 and f_2 are combined so that they produce a combination with a rank-score function that is more similar to f_b of Fig. 3.3, and Fig. 3.5 then this will minimize the false positives (FP) and false negatives (FN) associated with the combination with respect to f_b.

A concave-up rank-score graph, such as f_{b1}, assigns fewer ranks to the top scoring tracks and many to the lower scoring tracks, whereas a concave-down rank-score graph, such as f_{b2}, assigns many ranks to the top scoring tracks and few to the lower scoring tracks. Hsu and Lyons [9] refer to concave-up and down members of this family as complementary graphs. The rank-score graphs for the two scoring systems shown in Fig. 3.4 lead to the rank-score graphs shown as f_{b1} and f_{b2} in Fig. 3.5, and these are complementary rank-score graphs.

In general, two rank-score graphs won't be perfectly complementary as above, but if the rank-score graph of the combination is closer to the rank-score graph f_b of Fig. 3.3, then the FPs or FNs will be reduced. Details of this phenomenon can be found in Hsu and Lyons [9]. Hence in choosing a subset of features to fuse when tracking in crowded tracking scenarios, selecting features with complementary rank-score graphs will produce a result that minimizes false positives and false negatives.

Note that trackers with complementary rank-score graphs should be distinguished from trackers whose output is negatively correlated or independent. The latter is a relationship between the scores (i.e., the score function $s(d)$ for d in D, the set of all track hypotheses) the

trackers assign to a specific track. However, the former is a relationship between scoring behaviors (i.e., the rank-score function $f(i)$ for i in $N=\{1,2,\ldots,n\}$ and $|D|=n$), irrespective of the track being scored. Trackers may be correlated, negatively correlated or independent and still have complementary rank-score graphs. This gives our rank-score characteristic approach a distinctive advantage of characterizing the scoring behavior difference. It leads to a new approach to the quantitative and qualitative study of diversity among multiple scoring systems.

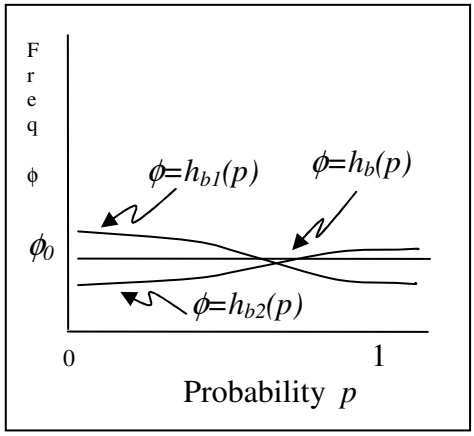


Figure 3.4: Histograms for two complementary scoring systems (h_{b1} and h_{b2}) in the crowded scenario.

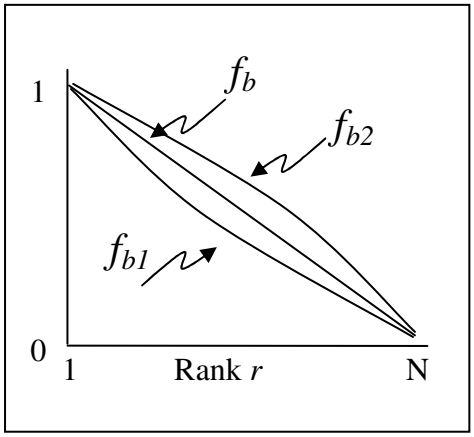


Figure 3.5: Rank-Score Graphs derived from histograms h_{b1} and h_{b2} respectively

4. Experiments

In this paper, we present two types of experimental results:

(1) **Type I:** These show that for crowded scenes, allowing a mix of score combination fusions and rank combination fusions can produce a significantly better tracking result. However, the experiments do not say how to choose operators to produce the improvement, only that the improvement is possible. These are in Section 4.1.

(2) **Type II:** These are the same as Type I except that we incorporate the rank-score graph information for selecting between fusions. They demonstrate that the difference of rank-score graphs criterion, proposed by Hsu and Taksa [11] as a diversity measure, is an effective way to select which fusion operation to perform. These are in Section 4.2.

We obtained ground truth information for twelve video sequences showing a variety of targets moving in indoor and outdoor scenes. The targets are not always easily separated from the background or each other, and in many sequences, they are close enough to each other to cause recurrent partial occlusions. The 12 video sequences used in the experiments were categorized in terms of whether they were indoor (shot in a large room) or outdoor (shot from a second story window looking onto a busy campus), how many targets were in the sequence, whether those targets crossed one another, and whether they were close enough to each other to cause recurring partial occlusions (overlapping, moving as a crowd) (see Lyons and Hsu [15] for more details). However, in each sequence, some targets can be separated most of the time, unlike the dense crowds studied in [21]. Ground truth was obtained by having a human observer go through the video sequence frame by frame and annotate the position of each target.

4.1 Mixed Combinations

In the first experiment, the RAF tracking system [10][16] was modified to carry out two fusion operations, a score fusion and an average rank fusion. In each case, the top m=30 tracks produced by tracker were evaluated against ground truth using a Mean Sum of Squared Distances (MSSD):

$$\frac{1}{nm}\sum_{j}\sum_{i}(gp_i - tp_{ij})^2$$

where gp_i $i=1..n$ is the ground truth sequence of target centroid image locations and tp_{ij}, $i=1..n$, is the jth best track's sequence of target centroid image locations. In addition, the top 30 tracks were examined to see which fusion operators had been used. The implementation and results of this experiment are presented and discussed in Lyons and Hsu [15] as RUN1 and RUN2 and not repeated here for brevity.

4.2 Selection of Combination using the Rank-Score Characteristics

Hsu and Taksa [11] have described how the difference in rank-score graphs can be used to guide whether a sum of score fusion or rank fusion will produce a better result. We will employ a simple measure of the difference as the variation of the rank-score functions. We define the difference between two rank-score functions f_A and f_B as follows for trackers A and B:

$$d(f_A, f_B) = \sum_{i \in N}(f_A(i) - f_B(i))$$

In our implementation we have three features. Let f_t be the rank-score graph for tracker t. We use the largest absolute difference between any two of the three features for selecting fusions:

$$\delta_{rs} = MAX \mid d(f_{t1}, f_{t2}) \mid for\ t1 \neq t2$$

Seq.	**RUN2** Score fusion		**RUN3** Score and rank fusion using ground truth to select		**RUN4** Score and rank fusion using rank-score function to select	
	MSSD Avg.	**MSSD Var.**	**MSSD Avg.**	**MSSD Var.**	**MSSD Avg.**	**MSSD Var.**
1	1537.22	694.47	1536.65	695.49	1536.9	694.24
2	816.53	8732.13	723.13	3512.19	723.09	3511.41
3	108.89	61.61	108.34	60.58	108.89	61.61
4	23.14	2.39	23.04	2.30	23.14	2.39
5	334.13	120.11	332.89	119.39	334.138	120.11
6	96.40	119.22	66.9	12.91	67.28	13.38
7	577.78	201.29	548.6	127.78	577.78	201.29
8	538.35	605.84	500.9	57.91	534.3	602.85
9	143.04	339.73	140.18	297.07	142.33	294.94
10	260.24	86.65	252.17	84.99	258.64	85.94
11	520.13	2991.17	440.98	2544.69	470.27	2791.62
12	1188.81	745.01	1188.81	745.01	1188.81	745.01

Table 4.1: MSSD Results for Type II Experiments
Lower MSSD implies better tracking performance.

The null hypothesis in our type II experiments is that this maximum difference of rank-score graphs is the same for fusion events where the score fusion produced the better results as for fusion events where the rank fusion produced the better result. If we disprove the null hypothesis, then this maximum difference is a useful criterion for selecting between fusion operations.

In the final phase of this experiment, we identify a threshold value for the maximum difference measurement using 8 of the 12 video sequences through the tracker but now using the maximum difference measurement (rather than the ground truth measurements) to select fusion operation. In addition, we run the tracker on 4 additional video sequences that were not used in the selection of the threshold operation. We compare the MSSD results obtained this way with those from the first experiment.

The ground-truth guided combination of score fusion and rank fusion (RUN3) of the type I experiments) was repeated, and the average and variance of the maximum difference of rank-score graphs was calculated separately for score and rank fusions for the four video sequences for which RUN3 showed a significant improvement (sequence 2, and sequences 6-8) in the Type I experiments. The average value of the difference for the score fusion operator, $\delta_{rs} = 0.05$, was then selected as a threshold value for this second set of experiments. If the variation between the rank-score graphs is less than or equal to δ_{rs} then a score fusion is used, otherwise a rank fusion is applied. All 12 video sequences were run, and the MSSD performance figures collected. The results are shown in Table 4.1 labeled as RUN4.

5. Conclusion and Future Work

The results in this paper present a data-driven, combinatorial fusion analysis approach to the problem of multitarget video tracking with occlusion. Our CFA approach differs from other approaches to the problem of fusion for video tracking in that it is data-directed and makes no assumptions about the characteristics of the targets. Unlike statistical or physical modeling approaches, it does not try to model the non-linear relationship between video features and target states during repeated mutual occlusions. Unlike heuristic approaches, the combinatorial fusion approach, being data-driven, has a quantifiable performance. Our tracking framework considers each feature measurement to be a separate *scoring system* on the set of target track hypotheses, and scoring behavior was characterized by the rank-score function. Two fusion operations were considered, an average rank fusion, and a Mahalanobis score fusion. In this paper, we proposed a measure of diversity $d(f_A , f_B)$ between two scoring systems (cues, features or tracking systems) A and B which is equal to the sum of differences between the two rank-score functions $f_A(i)$ and $f_B(i)$ across all ranking orders i in N. The measure of performance we used was the mean sum of squared differences (MSSD) between a hypothesized track and the ground-truth for the track, as established by a human observer. When comparing the performance of two fusion operations, we look at the average MSSD produced by the top 30 track hypotheses for each fusion. The one with the lower result was considered the fusion with better performance.

Out study suggests several issues and directions for future work. These include:

(1) Evaluation of performance: The MSSD measurement is used in this paper to evaluate the performance of a scoring system. In general, given two scoring systems, A and B, we like to find a criterion (or criteria) to predict the improvement of the combined scoring system $C(A,B)$. In this regards, the combination $C(A,B)$ is seen as a positive case if the performance of C, $P(C)$, is greater than or equal to the performance of A and B (i.e., $P(C) \geq max\{P(A),P(B)\}$). Otherwise it is a negative case [26]. See [8] for more results.

(2) Measurement of diversity: The difference of the rank-score functions f_A and f_B of two scoring systems A and B was used in this paper to represent the scoring diversity between A and B. That is, $d(A, B) = d(f_A , f_B)$. We will explore the possibility of using the rank functions, r_A and r_B, or the score function, s_A and s_B, and their variances $d(r_A , r_B)$ or $d(s_A , s_B)$ as diversity measurements respectively. The diversity $d(A, B) = d(r_A , r_B)$ was used in the information

retrieval domain and $d(A, B) = d(f_A, f_B)$ in virtual screening and drug discovery [26] and protein structure prediction [13] (see also references in [7]).

(3) Diversity of rank-score function: In this paper, we applied CFA to each target at each frame, F_i, $i=1,2,\ldots, f$. Our performance results and comparisons were based on averaging the MSSD's over all the frames. We will, in future work, explore the diversity $d(A,B)$ between a pair of scoring systems (cues, features or systems) across all frames of a tracking sequence. This will have to be done off-line on stored video sequences. However, exploring diversity along this dimension might shed some light on the variation between different cues, features or tracking systems in the long run. Let $F = \{F_1, F_2, \ldots, F_f\}$ be the set of frames in a video sequence. Let A and B be two cues, features or systems in the set of scoring systems $C = \{C_1, C_2, \ldots, C_m\}$. The **diversity score function** defined on F, $s_{(A,B)} = \sum_{j \in N} | f_A(j) - f_B(j) |$, where j is in $N = \{1, 2, \ldots, n\}$, $n=|D|$ and $D = \{d_1, d_2, \ldots, d_n\}$ is the set of tracks, and f_A and f_B are the rank-score functions of the scoring systems A and B respectively. It would lead to the **diversity rank function** $r_{(A,B)}(F_i)$ if we sort $s_{(A,B)}(F_i)$ into descending order. The **diversity rank-score function** $f_{(A,B)}(F_i)$ *is*:

$$f_{(A,B)}(j) = (s_{(A,B)} \, o \, r_{(A,B)}^{-1})(j) = s_{(A,B)} (r_{(A,B)}^{-1}(j))$$

where j is in $F^* = \{1, 2, \ldots, f\}$. The diversity rank-score function was defined and studied in the CFA framework ([7], [13]). Even though this measurement has to be calculated off-line, on a stored sequence of frames, it allows the diversity between two features across all frames to be studied. It is frame independent and may be more accurate when used in subset selection among cues, features or scoring systems for combination and fusion.

8. References

[1] Bar-Shalom, Y. and Fortmann, T., *Tracking and Data Association*. (1988): Acad. Press.

[2] Checka, N., and Wilson, K., *Person Tracking Using Audio-Video Sensor Fusions.* Proceedings of the MIT Project Oxygen Workshop, 2002.

[3] Cox, I.J. and Hingorani, S.L. *An Efficient Implementation and Evaluation of Reid's Multiple Hypothesis Tracking Algorithm for Visual Tracking*. Int. Conf. on Pattern Recognition (1994) pp.437-442.

[4] Gavrila, D., *The Visual Analysis of Human Movement: A Survey.* Computer Vision and Image Understanding (1999) **73**(1): pp. 82-98.

[5] Haritaoglu, I., Harwood, D., and Davis, L. *W4: Who, When, Where, What: A Real-time System for Detecting and Tracking People*. 3rd Int. Conf. on Face and Gesture Recog. (1998) pp.877-892.

[6] Hu, W.; Tan, T.; Wang, L.; Maybank, S., *A Survey on Visual Surveillance of Object Motion and Behaviors* Systems, Man and Cybernetics, Part C, IEEE Transactions on ,Volume: 34 , Issue: 3 , Aug. (2004) pp.334 – 352.

[7] Hsu, D.F., Chung, Y.S., and Kristel, B.S.; Combinatorial Fusion Analysis: Methods and Practice of Combining Multiple Scoring Systems. In: (H.H. Hsu, editor) *Advanced Data Mining Technologies in Bioinformatics,* Idea Group Inc, (2006).

[8] Hsu, D.F., Lyons, D.M., and Ai, J., Combinatorial Fusion Criteria for Real-Time Tracking. To appear: IEEE 20th Int. Conf. on Advanced Information Networking and Applications. Vienna, Austria, April (2006).

[9] Hsu, D.F., and Lyons, D.M., *A Dynamic Pruning Strategy for Real-Time Tracking*. IEEE 19th Int. Conf. on Advanced Information Networking and Applications. Taipei, Taiwan,

March (2005) pp.117-124.

[10] Hsu, D.F., Lyons, D.M., Usandivaras, C., and Montero, F. *RAF: A Dynamic and Efficient Approach to Fusion for Multi-target Tracking in CCTV Surveillance*. IEEE Int. Conf. on Multisensor Fusion and Integration. Tokyo, Japan; (2003) pp.222-228.

[11] Hsu, D.F. and Taksa, I., Comparing rank and score combination methods for data fusion in information retrieval, *Information Retrieval 8(3), (2005) pp.449-480.*

[12] Kittler, J., and Alkoot, F., *Sum versus Vote Fusion in Multiple Classifier Systems.* IEEE Trans. on PAMI (2003) **25**(1): pp. 110-115.

[13] Lin, K.L., Lin, C.Y., Huang, C.D., Chang, H.M., Jang, C.Y., Kin, C.T., Tang, C.Y., and Hsu, F.; Methods of improving protein structure prediction based on HLA neural networks and combinatorial fusion analysis. *WSEAS Trans. on Information Science and Applications. 2(12) (2005), pp.2146-2153.*

[14] Loy, G., Fletcher, L., Apostoloff, N., and Zelinsky, A. *An Adaptive Fusion Architecture for Target Tracking*. Proceedings of the 5th Int. Conf. on Face and Gesture Recog. Washington DC (2002) pp.261-266.

[15] Lyons, D., and Hsu, D. F.; Methods of combining multiple scoring systems for target tracking using rank-score characteristics, submitted, 2006.

[16] Lyons, D., Hsu, D.F., Usandivaras, C., and Montero, F. *Experimental Results from Using a Rank and Fuse Approach for Multi-Target Tracking in CCTV Surveillance. IEEE Intr. Conf. on Advanced Video & Signal-Based Surveillance*. Miami, FL; (2003) pp.345-351.

[17] Nandhakumar, N., and Aggarwal, J.K., *Physics-based Integration of Multiple Sensing Modalities for Scene Interpretation.* Proc. of the IEEE. V85, N1, Jan. (1997). pp.147-163.

[18] Moore, J.R., and Blair, W.D., *Practical Aspects of Multisensor Tracking* in: Multitarget-Multisensor Tracking (Eds. Y. Bar-Shalom, W.D. Blair) Artech House (2000) pp.1-76.

[19] Rasmussen, C., and Hager, G., *Joint Probabalistic Techniques for Tracking Multi-Part Objects. Proc. Comp. Vision & Pattern Recognition*. Santa Barbara, CA; (1998) pp.16-21.

[20] Reid, D., *An Algorithm for Tracking Multiple Targets.* IEEE Trans. Aut. Ctrl. Vol. AC-24, No. 6. Dec. 1979. pp. 843-854.

[21] Reisman, P., Mano, O., Avidan, S., Shashua, A., Crowd Detection in Video Sequences. *Symp. on Intelligent Vehicles.*Parma, Italy; (2004) June 14-17. pp 66-71.

[22] Sharma, R.K., *Probabalistic Model-Based Multisensor Image Fusion*, Ph.D. Diss. 1999, Oregon Grad. Inst. Science & Tech.: Portland, OR.

[23] Snidaro, L., Foresti, G., Niu, R., and Varshney, P. *Sensor Fusion for Video Surveillance*. 7th Int. Conf. on Information Fusion. Stockholm Sweden, (2004) pp.739-746.

[24] Varshney, P.K., *Special Issue on Data Fusion.* Proc. IEEE (1997) **85**(1).

[25] Xu, L., Krzyzak, A., and Suen, C.Y., *Method of Combining Multiple Classifiers and their Application to Handwriting Recognition.* IEEE Trans. SMC (1992) **22**(3): pp. 418-435.

[26] Yang, J.M., Chen, Y.F., Shen, T.W., Kristal, B.S., and Hsu, D.F.; Consensus scoring criteria for improving enrichment in virtual screening. *J. of Chemical Information and Modeling* 45 (2005), pp 1134-1146.